DOCUMENT RESUME

ED 410 113 SE 060 507

AUTHOR Taylor, Catherine S.
TITLE An Investigation of Scoring Methods for Mathematics Performance-Based Assessments.
PUB DATE Mar 97
NOTE 42p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, March 24-28, 1997).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Elementary Secondary Education; Mathematical Concepts; *Mathematics Instruction; Mathematics Skills; *Performance Based Assessment; *Scoring
IDENTIFIERS Alternative Assessment

ABSTRACT

Three mathematics scoring methods are being used or explored in large scale assessment programs: (1) item-by-item scoring; (2) holistic scoring; and (3) "trait" scoring. This study investigated all three methods of scoring on three mathematics performance-based assessments. Mathematics assessment tasks were selected from a pool of pilot tasks because they could be scored using all three methods. Results of the study suggest that holistic scoring and item-by-item scoring methods provide similar information while trait scores for conceptual understanding and mathematics communication tapped into different aspects of student performance. Implications for the validity of scoring methods now in use for performance-based mathematics assessments are discussed. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 410 113

An Investigation of Scoring Methods for Mathematics Performance-Based Assessments

Catherine S. Taylor
University of Washington

Running Head: Scoring Mathematics Performance Assessments

Paper presented at the annual meeting of the National Council on Measurement in Education, March 24-28, Chicago, Illinois

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
C.S. Taylor
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
X This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

Abstract

Three mathematics scoring methods are being used or explored in large scale assessment programs: item-by item scoring, holistic scoring, and "trait" scoring. This study investigated all three methods of scoring on three mathematics performance-based assessments. Mathematics assessment tasks were selected from a pool of pilot tasks because they could be scored using all three methods. Results of the study suggest that holistic scoring and item-by-item scoring methods provide similar information; however, trait score for conceptual understanding and mathematics communication tapped into different aspects of student performance. Implications for the validity of scoring methods now in use for performance-based mathematics assessments are discussed.

An Investigation of Scoring Methods for Mathematics Performance-Based Assessments

As more large-scale testing programs incorporate performance-based assessments, many researchers are investigating issues relevant to these "alternative" assessment formats. Much of this focus has been on the reliability of these assessments (e.g., Baxter, Shavelson, Goldman, & Pine, 1992; Fitzpatrick, Ercikan, & Ferrara, 1992). To date there is a dearth of studies designed to assess the validity of different scoring methods except in writing. For writing assessment, the strategies used to look at the validity of scoring methods have included (a) correlations between students' writing and multiple-choice assessments of language expression and mechanics, (b) correlations between holistic scores for an extended written piece and analytic scores on short paragraphs designed to measure a single aspect of writing, (c) correlations between primary trait scores (e.g., scores for quality of persuasion, story narrative, or information presentation) and general impression holistic scores, and (d) factor analyses of various methods of scoring and measuring writing (see Miller & Crocker, 1990, for a review of these studies).

If performance-based assessments in areas other than writing are to be included in large-scale assessment programs, then the validity of scoring methods must be investigated. According to validity theory (e.g., Messick, 1989), the validity of score interpretation and use depends on fidelity between the constructs (concepts and processes) that are being measured and the scores resulting from the test. Hence, the validity of a performance assessment depends on both the nature of the task presented to the student and the resulting scores used to evaluate students' responses.

The *Curriculum and Evaluation Standards for School Mathematics* (National Council of Teachers of Mathematics [NCTM], 1989) have established several different dimensions of mathematics which now guide the development of performance-based assessments: concepts and procedures in traditional areas of mathematics, as well as mathematical problem-solving, communication, and reasoning, and connections within and beyond mathematics. A great deal of work is being done to develop engaging tasks that elicit students' understanding of mathematics concepts and procedures and their skill in solving problems, communicating mathematical ideas, and reasoning mathematically in authentic contexts. Do our current methods of scoring accurately capture the

information available from students' performance on these tasks? Is the information provided by current scoring methods a valid reflection of the complex and interdependent nature of mathematics?

This study was an attempt to assess whether different methods of scoring tap into and appropriately represent these dimensions of mathematics. As the numbers of programs using performance-based assessments in state and national assessments increases, and as more classroom teachers attempt to use performance-based assessments, the issue of how to score these assessments is often locally defined. Decisions about how to score performances on large-scale assessments are typically made by two different groups: measurement specialists who have notions of reliability as their first priority for assessment and/or curriculum specialists who want to use assessment to impact teachers' conceptions of mathematics and, as a result, how mathematics is taught. Current state programs have used or explored three different types of scoring for performance-based assessments in mathematics and reading: item-by-item scoring, holistic scoring, and "trait" scoring.

Mathematics assessments for the state of Maryland are classified in terms of the Maryland educational outcomes and scored using rules that are specific to the item even when performance on an item depends on successful performance on related items (Fitzpatrick, Ercikan, & Ferrara, 1992; Yen, 1993a, 1993b). Mathematics assessments for the California Learning Assessment System (CLAS) were scored using holistic mathematics rubrics that were applied to multiple-step mathematics investigations, and open-ended mathematics problems (CLAS, 1993). The state of Washington is now exploring the use of what Arter (1993) calls "trait scoring" for mathematics assessments (Commission on Student Learning, 1996a). These traits represent different dimensions of performance and have been derived from the *NCTM Standards* (NCTM, 1989) and the state's Essential Learning Requirements document (Commission on Student Learning, 1996b). Trait rubrics are applied across multiple items linked to a single mathematics context or to open-ended mathematics problems.

Each of these three scoring methods for performance-based assessments has advantages and disadvantages. Item-by-item scoring results in more scores for examinees which results in higher estimates of score reliability. The problem is that, if items are linked, performance on some items may depend on performance on other items (Yen, 1993a). In addition, classifying an item as measuring

only a single aspect of mathematics (e.g., only measurement but not communication or reasoning), may be an artificial and arbitrary classification.

Holistic scoring involves the application of rubrics that represent conceptual and procedural understanding, mathematical communication, reasoning, and problem-solving in a single score. This may represent the complexity of mathematics and decrease scoring time; however, agreement among raters can be impacted by the idiosyncratic weights applied to dimensions of performance by different raters (Diederich, French, & Carlton, 1961; Freedman, 1979; Raforth & Rubin, 1984). Holistic methods also assume a single "trait," an assumption that is rarely supported by factor analytic research (Quellmalz, et al., 1982; Breland, Camp, Jones, Morris & Rock, 1987). Holistic scores provide a basis for comparison of examinees but do not give useful diagnostic information (Charney, 1984).

Trait scoring gives readers an opportunity to assign separate scores for two or more dimensions of mathematical proficiency (e.g., measurement, statistics, *and* communication) based on a single performance. This requires training of raters for each of the relevant dimensions of performance which some assessment specialists see as the "heart of the matter" (Arter, 1993). Teachers may have difficulty making distinctions between each of the dimensions of performance and scoring time may be doubled with the addition of a second trait, tripled with the addition of a third trait, etc. The purpose of this study was to investigate these three scoring methods when applied to the same tasks in order to answer two basic research questions: 1) Do all three methods result in equivalent score meaning? 2) Can locally dependent items or steps within a task be classified according to a single dimension of performance?

Methodology

The study was part of a pilot testing program for alternative assessments in the state of Washington. The purpose of the program was to create models for assessments to help guide the development of the upcoming state assessment system (which will include both external and classroom-based assessments). For the program as a whole, a total of 14 pilot mathematics tasks were piloted in middle/junior high school and high school. For this paper, three task models were selected from the pool for further investigation.

Instruments

This study focused on three mathematics tasks: two at the high school level (Number Cubes Game and School Dropout) and one at the middle/junior high school level (Garden Fence). The Number Cubes Game assesses applications probability theory to fair and unfair games. The School Dropout task assesses students' ability to read, interpret, and make predictions from graphs. The Garden Fence task assesses application of concepts of area and perimeter. These tasks were selected because they were carefully scaffolded (yielding several steps or items), had functioned well during the pilots, and could be scored using all three scoring methods (see Appendix A). Additional funding would have permitted the extension of this investigation to a larger number of tasks; however, these data permitted an initial investigation of scoring procedures.

Development of Assessment Tasks. The mathematics tasks were drafted by pairs of teachers. Tasks were edited by a professional test development staff and reviewed by all item/task writers. Tasks were then pre-piloted with 8 to 10 students to determine whether directions made sense. They were then revised based on the results of the pre-pilots and input from the reviewers. Finally, tasks were prepared for printing and distribution.

The task writers endeavored to create tasks that they believed would mirror processes used in classrooms as well as tasks that resembled more traditional classroom tests with open-ended items. The three mathematics tasks used in this study included three parts:

1. Setting a context: Students looked at stimulus materials that included written text, graphs, or tables that presented mathematical information.

2. Short Answer Items: Students responded to several short-answer items designed to have them analyze the information presented in the stimulus materials or extend their conceptual understanding to a new problem.

3. Integration Item: Students used the previous analyses to write an extended discussion, which could include predictions and/or conclusions, of the information presented in the task. This item also gave students an opportunity to bring closure to the task.

All tasks included introductory directions to tell students what steps they would be completing during extended tasks, as well as the bases for evaluation.

Scoring Methods. Three methods were used to score the mathematics tasks: holistic scoring, trait scoring, and item-by-item scoring. The holistic scoring rubrics were derived from the work done in the state of California (California Assessment Program [CAP], 1991, 1993). The 1993 CAP rubric was used as the basis of task specific six-point holistic rubrics for these three tasks. The CAP holistic rubric was modified to focus on the unique features of the given task. Figure 1 is the holistic rubric used for the Garden Fence task.

---

insert Figure 1 about here

For *trait* scoring, students received two scores: one for understanding of mathematical concepts and procedures relevant to the task and one for mathematical communication. The trait rubrics were developed by a committee of mathematics teachers from Washington state and refined using student work from the pilot assessment program during the spring and summer of 1995. The scoring rubrics were each on a four point scale. Figures 2A and 2B are the scoring rubrics for concepts and procedures of statistics and probability and for mathematical communication. These rubrics were applied to the Number Cubes Game and School Drop-Out tasks. The mathematical communication rubric (Figure 2B) and a measurement rubric were applied to the Garden Fence task. The concepts and procedures rubrics focused on understanding concepts and accurate application of appropriate procedures as defined for each conceptual area (statistics and probability or measurement). The communication rubric focused on clarity, organization, appropriateness, and completeness of communication. Raters applied each of the trait scoring rubrics to students' responses across the task as a whole.

---

insert Figures 2A and 2B about here

The third scoring method involved *item-by-item* scoring. Each of the items for a task was scored using a 2, 3, 4, or 5 point scoring rule depending on the complexity of the item. Scoring rules

for individual items took into account both accuracy of conceptual and procedural understanding, completeness of a response in addressing the requirements of the item, and, where appropriate, clarity of communication. The general schemes used to guide item specific scoring rule development were as follows:

1. Two-point rules - accurate (1) or not accurate (0).

2. Three-point rules - complete and accurate (2), partially complete and accurate or complete and partially accurate (1), largely incomplete, mostly inaccurate, or off task (0).

3. Four-point rules (for conceptual understanding) - accurate or reasonable, thoroughly addresses prompt (3), minor errors or gaps in logic but thoroughly addresses prompt (2), partially complete with significant errors or gaps in logic using appropriate procedures (1), largely incomplete, mostly inaccurate, or off task (0).

4. Four-point rules (for communication) - thoroughly and clearly communicates ideas (3), minor omissions but clearly communicates ideas (2), notable omissions or lack of clarity in presenting ideas (1), very difficult to understand (0).

5. Five point rules - accurate, effectively communicated, thoroughly addresses prompt (4), complete with minor errors, effectively communicated (3), partially complete, minor errors, clearly communicated (2), partially complete with significant errors using appropriate procedures, communication is acceptable or partially complete, accurate, but communicate is difficult to understand (1), largely incomplete, mostly inaccurate, or off task (0).

Item-by-item scoring rules were developed for each item or step in the task. Student work from a random sample of the papers was used to ensure that scoring rules were appropriate for the item and addressed as many contingencies as possible. Prior to being used in the scoring process, all scoring rules were reviewed by a mathematics editor from a professional test development company with extensive experience in large scale performance-based assessment programs. Minor modifications to scoring rules were made during training for scoring based on the input of the raters. An example of a scoring rule for one item of the Number Cube Game task is given in Figure 3.

insert Figure 3 about here

Procedure

Sample. In order to identify classrooms in which to pilot tasks, all of the state's 296 districts were invited to participate in the pilot program during the spring of 1995. District administrators were asked to volunteer 1 or 2 heterogeneously grouped classrooms at grades 7, 8, 10 and 11. Eighteen districts volunteered to participate in the high school pilots and twelve districts volunteered to participate in the middle/junior high school pilots. Once all pilot sites were identified, test forms were randomly assigned to classrooms in five different districts. Each test form contained a single mathematics task.

Materials were packaged for individual teachers. Materials included: a general overview of the pilot testing program with a description of the all eight task models being piloted along with specific directions for administering and returning the tasks, oral directions, sufficient student response books for one class of students, parent permission forms, student survey forms, and postage paid return envelopes. Teachers administered the tasks and returned materials in postage-paid return envelopes. All materials were received between March 31 and May 31, 1995.

Of the five classrooms that were sent materials for each task, student responses were returned from four districts for the Garden Fence task, three districts for the Number Cubes Game task, and all five districts for the School Dropout task. The number of students who completed the Garden Fence task were from grades seven (N = 15) and eight (N = 55). The number of students who completed the Number Cubes Game were from eleventh grade (N = 53). The number of students who completed the School Dropout task were from grades ten (N = 16) and eleven (N = 63).

Raters and Rater Training. Six research assistants were hired to score the mathematics tasks. Two raters were mathematics teachers and four were advanced masters students in mathematics education. Raters rotated through all three scoring procedures across the three different tasks. Raters were paired differently for each task. So, for a given task, one pair scored the tasks using item-by-item scoring procedures, one pair scored the tasks using two trait rubrics (e.g., understanding of statistics

and probability concepts and mathematics communication), and one pair scored the task using the holistic scoring rubric. For the next task, raters were placed in new pairs and trained to use a different type of scoring procedure. In that way, results across the tasks were not dependent on an interaction between score type and rater or some unique characteristic of a given pair of raters. Pairs of raters were trained separately so that they could focus on the scoring rules they were to apply to a given task. Raters were not aware of the purpose of the study, although they were certainly aware that they were applying different scoring methods to different tasks.

All raters participated in a 2 hour training session for each scoring method to be applied. Prior to beginning the scoring process, raters completed the relevant task themselves. They discussed the task and what was demanded by it. They were then given exemplary responses (prepared by the item/task writers) and discussed the similarities and differences between their own responses and those prepared by the teachers who wrote the tasks. Given the nature of the tasks, a single set of correct responses for each task was not possible, however, the sample answers indicated a range of possible acceptable responses.

The pair of raters who used item-by-item scoring rules reviewed and discussed the rules and then scored training sets of papers independently. (Training sets represented a range of student work. Training sets were selected and scored by the researcher and validated via scores assigned to the papers by two research assistants trained in scoring procedures during the summer of 1995.) Raters discussed their scoring decisions with the researcher and each other, and worked toward a consensus with the criterion scores. They scored another training set, reaching even closer agreement with the criterion scores (range of agreement was 78-85%), and then worked toward a consensus on scores. Then they scored the papers in the research set.

The pair of raters who applied the trait scoring rules reviewed the scoring criteria for the concepts and procedures trait first (measurement or statistics/probability) and discussed its meaning, including how the trait was distinct from other relevant dimensions of mathematics and how the scoring rules could be applied to make a judgment across all responses in the task. Raters then scored the first training set. Once they had completed scoring the papers, raters met with the researcher and

discussed their ratings. For all papers, the scores for each rater were within one point of the criterion scores. Raters discussed all papers, attending to points of agreement as well as disagreement, and worked toward consensus with the criterion scores given to each paper.

Raters then scored the second training set with greater fidelity to the criterion scores (exact agreement ranged from 70-80%) and repeated the discussion process. Raters scored the research set independently. Once raters completed scoring for the concepts and procedures trait relevant to the task, they repeated the entire process for the mathematical communication trait.

The pair of raters who applied the holistic scoring rubric reviewed the rubric and discussed its meaning, including the importance of considering all relevant dimensions of mathematics in their judgments and how a single rule could be applied to make a judgment across all responses in the task. Raters then scored the first training set. Once they had completed scoring the papers, raters met with the researcher and discussed their ratings. For all papers, the scores for each rater were within one point of criterion scores. Raters discussed all papers, attending to points of agreement as well as disagreement, and worked toward consensus with the criterion scores assigned to each paper. Raters then scored the second training set with greater fidelity to the criterion scores (exact agreement ranged from 80-86%) and repeated the discussion process. Raters scored the research set independently.

Exact agreement among raters for each of the scoring methods and for each of the tasks in the research set was within expected ranges for performance based assessments. Exact agreement for the item-by-item scoring method ranged from 65-100% across the items in the three tasks. Exact agreement for the holistic scoring method ranged from 78-80% across the three tasks. Exact agreement between raters for the trait scoring method ranged from 76-84% across the two traits and the three tasks. A research assistant who had not participated in the scoring recorded the scores and identified the papers with discrepant scores. The primary researcher scored all papers that received discrepant scores. Final scores assigned to tasks with discrepant scores were based on the agreement between the primary researcher and one of the two raters.

At the end of the data gathering process, raters were asked to discuss the three scoring methods in terms of their preferences and the strengths and weaknesses of each method from their perspectives

as teachers. Their discussion was facilitated by the primary researcher. Comments that bear on interpretation of the data analysis results are presented in the discussion that follows.

Data analysis

Data analyses for this study posed two problems: 1) there was no "absolute" in terms of the true score against which to judge all other scores and 2) different students completed different tasks. The first problem addressed in data analysis was in terms of what would be considered the "true" score. "It is necessary to remember that a true score. . . is a theoretical idea. This score will not completely reflect the 'true' characteristic of interest unless the test has perfect validity—that is, unless the test measures exactly what it purports to measure" (Allen & Yen, 1979, p. 60). A question posed by this study is, "Do all three methods result in equivalent score meaning?" Hence, each method can be considered one way to define the examinee's true score. Once scores are obtained for each examinee for each scoring method (using rater consistency as the vehicle for determining the reliability of the score), each type of score could be used as the "criterion" against which the other score methods are compared. In the following analyses, all three score types served as criterion scores for at least one analysis. The second problem was in how to deal with the independence of subjects and tasks. The decision was to approach the analyses as is done in mathematical investigations wherein one looks across cases to determine whether there are patterns that can be discerned in the data.

Choices of analyses were based on strategies that have been used to analyze scoring methods for writing assessment (e.g., Breland, et al. 1987; Diederich, et al. 1961; Quellmalz, et al. 1982). Each of these analyses served a different purpose in this investigation. Analyses included:

1. Descriptive information about the scores for each task. Descriptive data demonstrates whether each scoring method makes a similar statement about the "typical" performance of examinees. Hence, a look at task means and standard deviations can indicate whether all methods are consistent in terms of describing the difficulty of the task.

2. Correlations between the sum of item scores for each task, trait scores, and holistic scores. Correlations between the sum of item scores, trait scores, and holistic scores were obtained in order to see whether these scoring methods were providing similar information about examinees. Item

scores were summed for each task because that is the typical method used to obtain a test or task score. If the three score types are equally valid, then correlations between them should be high and the shared variance should be substantial.

3. Regression analyses using holistic scores as the criterion measure for either item scores or trait scores and regression analyses using each trait score as the criterion measure for item scores. Regression analyses were used to determine (a) the amount of variance of each criterion measure explained by the predictor scores, (b) which items were the best predictors of the holistic or trait scores, and (c) whether both traits were useful in predicting of the holistic scores. Two types of regression analyses were conducted. Simultaneous regressions (entering all item scores or trait scores in the prediction of holistic scores) and step-wise regressions for each criterion/predictor relationship being investigated. Simultaneous regressions were useful in determining how much of the criterion score variance was explained by the entire set of predictors. Step-wise regressions were useful in determining which predictors were the *best* predictors of the criterion scores. In order for a variable to enter the step-wise regressions process, it had to contribute significantly ($p < .05$) to the increase in R^2.

4. Factor analyses of the item scores. Factor analyses of item scores were conducted in order to assess whether the items within a task were all measuring the one factor or multiple factors. If multiple factors, the question was whether items designated as "concepts and procedures" items loaded on a separate factor from items designed as "mathematical communication" items. Results of these analyses follow.

Results

Descriptive data. Tables 1 through 3 present the means and standard deviations for the holistic, trait, and sum of item scores for each task as well as the correlations between scores. Using item scores, there were fifteen total points possible for the Garden Fence task. Students tended to perform somewhat poorly (mean = 5.84) over the set of items using item-by-item scoring. Similarly, typical scores on the holistic and trait rubrics were fairly low. The mean holistic score was 2.49 out of 6 levels; the means for the communication and measurement scores were 1.69 and 1.77 out of 4 points

respectively. There were fourteen points possible for the Number Cubes Game task. Students did fairly well using item-by-item scoring with a mean score of 10.79. Similarly, students did moderately well on the holistic and trait rubrics. The mean holistic score was 3.98 out of 6 points; the means for the communication and probability and statistics scores were 2.44 and 2.54 out of 4 points respectively. Finally, there were fifteen points possible for the School Dropout task. Students did fairly well using item-by-item scoring with a mean score of 11.25. This result was somewhat inconsistent with the moderately low performance as judged through holistic and trait scoring methods. The mean holistic score was 3.17 out of 6 points; the means for the communication and probability and statistics scores were 2.22 and 2.15 out of 4 points respectively.

---

insert Tables 1 through 3 about here

Correlations between the sum of item scores for each task, trait scores, and holistic scores. For two of the tasks (Garden Fence and Number Cube Game), correlations between the mathematical trait scores were lower than correlations between the trait scores and holistic scores or between the sum of item scores and the holistic or trait scores. This would suggest that the trait scores are measuring somewhat distinct traits and that the traits are more strongly associated with overall mathematical power than with each other. However, for the third task (School Dropout) the reverse was true; the correlation between trait scores was higher than all other correlations.

For all three tasks, the correlations between the sum of the item scores and the holistic scores (.705 to .910) were higher than correlations between trait scores and holistic scores (.552 to .848) . This supports evidence found in the regression analyses suggesting that there is a stronger relationship between holistic scores and item scores than between holistic scores and trait scores. In addition, for all three tasks, the correlations between the sum of item scores and the concept and procedures trait scores were slightly higher (.746 to .815) than the correlations between the sum of the item scores and the communication scores (.661 to .813).

Regression analyses using holistic scores as the criterion measure for either item scores or trait scores. The results of the regression analyses for holistic scores suggested that item scores explained

more of the holistic score variance than did trait scores when all variables were entered simultaneously. Table 4 shows the amount of holistic score variance explained by the set of items or the two trait scores for each task. Variance of holistic scores explained (adjusted R^2) ranged from .469 to .846 for item scores and from .410 to .797 for trait scores.

---

insert Table 4 about here

When stepwise regression analyses were performed using the item level scores (see Table 5), several patterns emerged. To begin with, the first item to enter the equation for every task was one that required significant reasoning about the problems presented by the task. For example, the first item to enter the equation for the Garden Fence task was the one in which students had to explain and show whether or not a larger garden could be obtained from the same fence (Isaac's idea). The first item to enter the equation for the Number Cubes Game task was one in which students had to write a letter to the game company and explain and show whether the original game was fair and describe the game they had developed that was a fair game. Finally, the first item to enter the equation for the School Dropout task was one in which students had to explain whether it was possible for school dropout numbers to decrease while the dropout rate was increasing. In each case, the item goes beyond application of a simple algorithm or procedure.

---

insert Table 5 about here

A second pattern was that for all three tasks, using a $p < .05$ level of significance, most items were useful in explaining holistic score variance. For the Garden Fence task, four out of five items significantly contributed to the overall variance explained. For the Number Cube Game task, three out of five items entered the equation. For the School Dropout task, five out of seven items entered the equation. In each case, the items that did not enter the equations assessed understandings that students also showed in other items. For two tasks (Garden Fence and Number Cube Game), the items that did not enter were ones in which students drew conclusions based on what they had learned from doing one or more steps in the task. Conclusions depended on the success students had completing the

relevant investigation. For the third task (School Dropout), one item required students to describe the enrollment trend given in the data and the other item asked students which graph supported a position that the school dropout problem was getting better. Both of these items were answered correctly by 78 out of 79 students in the sample.

For the step-wise regressions using trait scores as predictors and the holistic scores as the criterion, both traits entered the equation for the Garden Fence task and for the Number Cube Game task. Only the statistics and probability score entered the equation for the School Dropout task suggesting that holistic scores were largely a function of conceptual and procedural understanding for that task.

Regression analyses using each trait score as the criterion measure for item scores. The results of the regression analyses using item scores to explain trait scores suggested that item scores were generally less useful in explaining the trait score variance than they were in explaining holistic score variance. Table 6 shows the amount of trait score variance explained by the set of items for each task. Variance of communication scores explained by item scores ranged from .433 to .661 and variance of concept and procedure scores explained by item scores ranged from .518 to .661.

---

insert Table 6 about here

When step-wise regressions were conducted to see which items were most useful in explaining the variance of the trait scores, it is clear that different items figured into the variances for each trait (See Table 7). For all tasks, items that were useful in predicting the communication scores were those that required communication about fairly simple conceptual understandings or a presentation of conclusions in a letter form. For all tasks, the items that were useful in predicting the concept and procedures trait scores were those that required applications of conceptual understandings. In each case, however, there was an overlap of items.

---

insert Table 7 about here

Factor analyses of item scores. A factor analysis was conducted for each task. The purpose was to investigate whether items clustered as they were expected to (i.e., items intended to assess mathematical communication were expected to load on one factor and items intended to measure the conceptual understandings were expected to load on a second factor). Since this was an investigation, exploratory rather than confirmatory factor analysis procedures were used using SPSS principal components analysis. An orthogonal rotation procedure was used, although there were no differences in the items loading on factors when various oblique rotation procedures were used.

The factors that emerged from the factor analyses across the three tasks turned about to be very consistent. For all three tasks, one factor (henceforth called the "higher order thinking skills" or HOTS factor) was composed of items that required students to generalize fairly standard mathematical procedures to new situations, to make predictions from trends, to communicate mathematical predictions or conclusions, or to explain one or more mathematical relationships. In the Garden fence task, four items loaded on the HOTS factor: three of these items asked students to demonstrate and explain whether they could use more fence, the same fence, and less fence to obtain a larger garden; the fourth asked them to decide which solution was the best one and why. In the Number Cubes Game task, two items loaded on this HOTS factor: one asked students to generate a new, fair game by changing numbers on the number cubes and describe why it was a fair game; the other asked students to write a letter to the game company explaining and showing the fairness of the original game and the new game. Finally, in the school drop-out task, the HOTS factor was composed of two items: one that asked whether it was possible for dropout numbers to decrease while dropout rate increased and a second that asked students to write a letter to the governor predicting school dropout in 1995 based on the trends in the data. Finally, across the three tasks, the variance associated with the HOTS factor ranged from about 25% for the School Dropout Task to about 60% for the Garden Fence task.

For each task, one factor included one or more items that were fairly simple applications of mathematical procedures (henceforth called the PROCEDURE factor). For the Garden Fence task, the PROCEDURE factor included only the item that asked students to compute the area and perimeter of the original garden. For the Number Cubes Game task, the PROCEDURE factor included three items:

one item asked students to create a display showing the outcomes of the original game, one item asked whether the original game was fair to both players, and one item asked students to compare the prediction they made prior to investigating the game with what they believed after investigating the game. Finally, for the School Dropout task, two items loaded on the PROCEDURE factor: each item asked students to describe a trend (dropout numbers and enrollment numbers) in the data, using data to support their claims about the trend. Finally, the variance associated with the PROCEDURE factor ranged from about 14% for the Garden Fence task to about 16% for the School Dropout task.

The third factor for the School Dropout task is somewhat interesting. Two items loaded on this factor. For each item, students were given a position about the trends in school dropout. Students had to decide which data supported the proposed trends. These items required use of a slightly different thinking strategy than simply reporting data. This factor might be called a SUPPORT factor since students were asked to support a conjecture but were not asked to make the conjecture themselves. The item that did ask them to make their own conjectures loaded on the HOTS factor. The variance associated with the SUPPORT factor was about 19%.

One final factor emerged for the School Dropout task that included a single item: dropout rate. In this item, students read data and a graph that showed that dropout rate was steadily increasing by about 1% per year. It was different from the other PROCEDURE items only in that it required students to understand percent changes rather than number changes.

Given the results of these factor analyses, it appears that factors were based more on the type of thinking required of the students in given items than on distinctions between the traits theoretically measured by the task. This supports the information from the regression analyses suggesting that some items measure more than one trait.

Discussion and Conclusion

This study was an investigation of three scoring methods for performance-based mathematics assessments. While it is clear that these data cannot be used to support claims for the most *appropriate* method of scoring performance tasks, the investigation provides an initial look at how these methods may relate and what makes each unique. The questions addressed by the study were: 1) Do all three

methods result in equivalent score meaning? 2) Can locally dependent items or steps within a task be classified according to a single dimension of performance?

Do all three methods result in equivalent score meaning? The data suggest that different scoring methods tap into different elements of students' performances. For example, in the factor analyses, the majority of items within a task were useful in explaining the variance for the holistic score. This suggests that holistic scoring methods reflect, to a large extent, a "sum of item scores." It may also suggest that task specific rubrics and item specific scoring rules, especially when developed by the same test developer, largely focus on the same elements of students' responses. When trait scores were used as the criterion variables, smaller and overlapping sets of items were useful in explaining trait score variance, suggesting that the two trait scores tapped unique aspects of students' responses, although some items related to both conceptual and procedural understandings and mathematical communication. For all three tasks, trait scores had lower correlations with holistic scores than did the sums of item scores. This result also suggest that when raters applied item level scoring and holistic scoring they looked at similar elements of students' work. Again, this may be due to the fact that the holistic rubric and item-specific rules were adapted to requirements of each task.

Another conclusion that can be drawn is that trait scores, while sharing variance (52-64%) do seem to tap into unique aspects of students' work. This could support the use of trait scoring if holistic scoring methods are seen as more desirable than item-by-item scoring procedures. The use of trait scores could provide diagnostic information at the standards level and still allow for holistic judgments about students' depth of understanding and skill.

Can locally dependent items or steps within a task be classified according to a single dimension of performance? These data suggest that constructs, such as mathematical communication, cannot be isolated from the conceptual context. For all three tasks, one or more items were useful in explaining variance for both the concepts and procedures trait and for the communication trait. This suggests that test development strategies that attempt to classify items as measuring *either* mathematical concepts *or* mathematical communication (or reasoning or problem-solving) may not be valid. More research is needed in this area.

One of the problems with this study is that not all of the tasks were equally useful in providing information that was helpful to the study. The School Dropout task was not a strong task. Using the item-by-item scoring method, the general performance on the task as a whole was quite high (mean = 11.25 out of 15 points). The percent of students receiving perfect scores for the five of seven items was quite high (84% to 100%). Most of the task variance depended on two items: one in which students were asked whether it was possible for drop-out numbers to decrease while drop-out rate increased and one in which students wrote letters to the governor predicting trends and explaining their predictions. In addition, the three scoring methods resulted in fairly different overall judgments about students. While both the holistic and trait score judgments were that the students did somewhat poorly on the task, the sum of the item-by-item scores suggested that students did quite well. One benefit of this inconsistency is that it shows that raters differentially weight steps within a task. When the raters evaluated the students' responses to the task as a whole, they seem to have given greater weight to the two, more difficult items. This would make sense because these were the HOTS items. These two items demanded more thinking and a better understanding of statistics and probability than did the five items that required fairly rote responses.

A second problem with the study was in the number of cases for each task. Both the pool of volunteer districts was small and not all of the teachers who volunteered returned completed tasks. Future studies are needed with more students per task; however, given the costs associated with scoring, exploratory studies are also needed.

The investigation presented here generates a number of possible research questions. One question is: "What factors influence raters' applications of trait and holistic scoring rubrics?" The regression analyses presented here suggest that when raters use scoring rubrics that measure different traits, they look at different elements of student performance. It also suggests that the use of task specific holistic scoring rules results in much the same information as do sums of item scores. This is a result that merits more research to see what teachers in less controlled settings think about when applying trait rubrics. Having raters talk aloud during scoring sessions would help us capture the reasoning used by raters as they score tasks.

A second question is: "What influence do methods of scoring have on teachers' conceptions of the discipline of mathematics?" If raters can be taught to look at different elements of students' work to apply trait rubrics, teachers can also be taught to do so. Would this form of scoring help teachers develop a broader conception of mathematics than do our current methods of scoring? Would methods of scoring that attend to multiple dimensions of student performance influence how teachers teach mathematics?

A third research question is" "Would these results occur if the traits scored included reasoning and/or problem-solving?" The three tasks in this study were scored only for conceptual understanding and mathematical communication. The focus on communication was chosen because all three tasks demanded mathematical communication skills. It would be useful to look at several tasks that ask students to explain and show their reasoning or several tasks that focus on problem-solving processes. We need to know whether raters focus on conceptual and procedural accuracy when assessing problem-solving and reasoning or whether teachers are able to make distinctions between these aspects of students' work.

Finally, and most important, what *is* the most valid way to score students' responses to complex, multi-step tasks. At this time, there is inadequate evidence to support the use of one method of scoring over any other; however, it is time for test developers take a serious look at evidence for the validity of different scoring methods. This investigation is a beginning look at three methods of scoring. Studies are needed that include more tasks, more raters, and more students to see whether the patterns that emerged in this investigation are replicated. Studies are needed that focus on subtle variations on these three methods. For example, how would the results have been different if the raters had used general impression holistic rubrics rather than task specific rubrics? How would the results have been different if the raters had used task specific trait scoring rubrics? Studies are needed that look at a wider range of assessments for single examinees to see whether each method of scoring is equally predictive of students' overall performance in mathematics.

Methods of scoring have been developed to serve different purposes. Their use can also lead to a variety of consequences - intended or unintended. A test composed of disconnected items may be

easier to score and may yield better reliability coefficients, but it can reinforce a notion that mathematics is a collection of discrete skills. This conception may be strengthened by state level programs that attempt to classify items in complex tasks for purposes of scoring and reporting. This study suggests that such classifications may be somewhat arbitrary. To the extent that arbitrary classifications lead to scores that do not truly reflect the demands of items and tasks, the validity of scores are threatened.

Holistic scoring rules were invented to acknowledge the complex nature of mathematics by integrating problem solving, communication, and conceptual understanding into one rubric. Based on this study, however, it appears that raters using task specific holistic rubrics attend more to conceptual understanding than to mathematical communication. Holistic scores also are limited in terms of helping diagnose students' strengths and weakness for purposes of instructional interventions.

The use of trait rubrics may improve the validity of the assessment of mathematical performances by giving raters (and teachers) an opportunity to assess the multiple dimensions of tasks while still reflecting the integration of different dimensions of mathematics required in mathematics problems. These data would suggest that trait scoring is a viable option for mathematics performance assessments. Still, trait rubrics, as used in this study, require multiple passes at scoring and raters must have a deep understanding of mathematics in order to use them. If future studies also support their use, they should be considered along with other scoring options. They may achieve a balance in the assessment of mathematical content and processes; providing the diagnostic information that some try to obtain through item-by-item scoring as well as the judgments of quality that are possible through holistic scoring.

Our scoring choices are part of what must be considered when we examine the validity of our assessments. We must consider the constructs (content and processes) we intend to measure and develop then scoring methods that are consonant with those constructs. While trait scoring may not work well with traditional psychometric conceptions that assume measured traits are independent of one another, multiple trait rubrics applied to performance tasks may reflect more of what is true (valid) in the constructs than what test developers wish were true.

References

Arter, A. (1993, April). *Designing scoring rubrics for performance assessments: The heart of the matter.* Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Baxter, G., Shavelson, R. J., Goldman, S. R., and Pine, J. (1992). Evaluation of procedure-based scoring for hands-on science assessment. *Journal of Educational Measurement, 29*, 1-17.

Breland, H. M., Camp, R., Jones, R. J., Morris, M. M., & Rock, D. A. (1987). *Assessing writing skill.* New York: College Entrance Examination Board.

California Assessment Program (1991). *A Sampler of Mathematics Assessment:*. Sacramento, CA: California Department of Education.

California Assessment Program (1993). *A Sampler of Mathematics Assessment: Addendum.* Sacramento, CA: California Department of Education.

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical overview. *Research in the Teaching of English, 18* (1), 65-81.

Commission on Student Learning for the State of Washington (1996). Making the connection: High school mathematics. Classroom activities connected to essential academic learning requirements. Olympia, WA: Author.

Diederich, P. B., French, J. W., & Carlton, S. T. (1961). *Factors in judgments of writing ability* (Research Bulletin 61-15). Princeton, NJ: Educational Testing Service. (ERIC Document Reproduction No. ED 002 172).

Fitzpatrick, A. R., Ercikan, K., & Ferrara, S. (1992, April). An analysis of the technical characteristics of scoring rules for constructed-response items. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Freedman, S. W. (1979). How characteristics of student essays influence teachers' evaluation. *Journal of Educational Psychology, 71*, 328-338.

CTB Macmillan/McGraw-Hill (1991). *Final Technical Report for the Maryland School Performance Assessment Program.* Monterey, CA: Author

Messick, S. (1989). Validity. In, *Educational Measurement* (Robert Linn, Ed.). Washington, DC: U. S. Department of Education.

Miller, M. D., & Crocker, L. (1990). Validation methods for direct writing assessment. *Applied Measurement in Education, 3*, 285-296.

National Council for Teachers of Mathematics (1989). *Curriculum and Evaluation Standards for School Mathematics*. Reston, VA: Author.

Quellmalz, E. S., Capell, F. J., & Chou, C. P. (1982). Effects of discourse and response mode on the measurement of writing competence. *Journal of Educational Measurement, 19*, 241-258.

Raforth, B. A., & Rubin, D. L. (1984). The impact of content and mechanics on judgments of writing quality. *Written Communication, 1*, 446-458.

Yen, W. (1993a). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187-213.

Yen, W. (1993b, June). *The Maryland School Performance Assessment Program: Performance assessment with psychometric quality suitable for high stakes usage*. Paper presented at the annual Large Scale Assessment Conference, Albuquerque, NM.

Table 1

Descriptive data for and intercorrelations between holistic scores, trait scores, and sum of item scores for Garden Fence task

	No. of Cases	Points Possible	Mean	SD	Holistic Score	Communication	Measurement	Sum of Item Scores
Holistic Score	69	6	2.49	1.45	1.00	.853 (.728)*	.848 (.719)	.910 (.828)
Communication	67	4	1.69	.76		1.00	.801 (.640)	.813 (.660)
Measurement	62	4	1.77	.84			1.00	.815 (.664)
Sum of Item Scores	69	15	5.84	3.24				1.00

* Common variance in parentheses

Table 2

Descriptive data for and intercorrelations between holistic scores, trait scores, and sum of item scores for Number Cube Game task

	No. of Cases	Points Possible	Mean	SD	Holistic Score	Communication	Statistics & Probability	Sum of Item Scores
Holistic Score	53	6	3.98	1.29	1.00	.730 (.533)*	.760 (.578)	.838 (.702)
Communication	52	4	2.44	.94		1.00	.721 (.520)	.738 (.545)
Statistics & Probability	52	4	2.54	.92			1.00	.746 (.557)
Sum of Item Scores	53	14	10.79	2.92				1.00

* Common variance in parentheses

Table 3

Descriptive data for and intercorrelations between holistic scores, trait scores, and sum of item scores for School Dropout task

	No. of Cases		Mean	SD	Holistic Score	Communication	Statistics & Probability	Sum of Item Scores
Holistic Score	79	6	3.17	1.08	1.00	.552 (.305)*	.641 (.411)	.705 (.497)
Communication	79	4	2.22	.64		1.00	.734 (.539)	.661 (.437)
Statistics & Probability	79	4	2.15	.72			1.00	.750 (.563)
Sum of Item Scores	79	15	11.25	2.09				1.00

* Common variance in parentheses

Table 4

Variance of holistic (mathematics power) score explained by item scores or trait scores for each mathematics task

Task	Predictor Variables	Multiple R	R^2	Adjusted R^2
Garden Fence Task	Item Scores	.926	.857	.846
	Trait Scores	.896	.804	.797
Number Cubes Game Task	Item Scores	.858	.736	.708
	Trait Scores	.804	.646	.631
School Dropout Task	Item Scores	.721	.519	.469
	Trait Scores	.652	.425	.410

Table 5

Items That Contribute Significantly to Variance of Holistic Scores for Each Task

Task	Predictor Variables	Multiple R	R^2	Adjusted R^2
Garden Fence	Same fence; more area	.922	.851	.841
	Less fence; more area			
	More fence; more area			
	Original Garden			
Number Cube Game	Letter to Game Company	.857	.734	.718
	Original (Unfair) Game			
	New (Fair) Game			
School Dropout	Possible?	.723	.523	.488
	Dropout Number			
	Dropout Rate			
	Problem is Better			
	Letter to Governor			

Table 6

Variance of trait scores explained by item scores for each mathematics task

Task	Criterion Variable	Multiple R	R^2	Adjusted R^2
Garden Fence Task	Communication	.828	.686	.661
	Measurement	.830	.689	.661
Number Cubes Game Task	Communication	.766	.586	.541
	Statistics & Probability	.779	.606	.563
School Dropout Task	Communication	.697	.487	.433
	Statistics & Probability	.750	.564	.518

Table 7

Items That Contribute Significantly to Variance of Trait Scores for Each Task

Task	Criterion	Predictors	Multiple R	R^2	Adjusted R^2
Garden Fence Task	Communication	Same fence; more area More fence; more area	.813	.662	.651
	Measurement	Same fence; more area Less fence; more area	.823	.678	.667
Number Cube Game Task	Communication	Letter to Company New (Fair) Game	.727	.528	.509
	Statistics & Probability	New (Fair) Game Original (Unfair) Game	.751	.563	.546
School Dropout Task	Communication	Letter to Governor Dropout Rate Possible?	.635	.403	.378
	Statistics & Probability	Letter to Governor Possible? Dropout Numbers Problem is Worse	.733	.537	.511

Level 6 The student's response shows a clear understanding of the requirements of the task. The response completely and thoroughly addresses the task and the concepts associated with it.

- Consistent use of appropriate algorithms to obtain area and perimeter.
- Accurate computations.
- Clear understanding of the relationship between area and perimeter both within and beyond the problem.
- Clear and appropriate diagrams for each of the proposed solutions are provided.
- Accurate and complete labels for all diagrams and results are given.
- Explanation for final choice is logical based on results of each investigation.
- Explanation for final choice clearly communicates understanding of the relationship between area and perimeter, as well as the proposed solution

Level 5 The student's response shows a very good understanding of the requirements of the task. The response completely addresses the task and the concepts associated with it.

- Consistent use of appropriate algorithms to obtain area and perimeter.
- Accurate computation.
- Clear understanding of the relationship between area and perimeter within the given problem.
- Diagrams for at least two of the proposed solutions are given.
- Labels for diagrams and results are mostly complete and accurate.
- Explanation for final choice is logical based on results of each investigation.
- Explanation for final choice clearly communicates understanding of the relationship between area and perimeter, as well as the proposed solution

Level 4 The student's response shows a good understanding of the requirements of the task. The response addresses the task and the concepts associated with it.

- Use of appropriate algorithms to obtain area and perimeter.
- Computations may include minor errors.
- Understanding of the relationship between area and perimeter for original perimeter but may have difficulty extending solution to smaller perimeter.
- Mostly understands how dimensions can be shifted to obtain different areas for each proposed solution.
- Labels for diagrams, if given, and/or results are mostly complete and accurate.
- Explanation for final choice is logical based on results of each investigation.
- Explanation for final choice shows understanding of the relationship between area and perimeter, as well as the proposed solution

Figure 1

Holistic Scoring Rubric for Garden Fence Task

Level 3 The student's response shows a fair understanding of the requirements of the task. The response mostly addresses the task and the concepts associated with it.

- Use of appropriate algorithms to obtain area and perimeter.
- Computations may include several errors.
- Understanding of area and perimeter but may have difficulty extending the solution.
- Diagrams and labels may be omitted
- Explanation for final choice is logical based on results of each investigation.
- Explanation for final choice communicates some understanding of the relationship between area and perimeter

Level 2 The student's response shows a poor understanding of the requirements of the task. The response addresses some components of the task and the concepts associated with it.

- Use of appropriate algorithm to obtain area OR perimeter.
- Computations may include errors that detract from solutions.
- Weak understanding of the relationship between area and perimeter.
- Accurate labels may or may not be given for diagrams and results.
- Final choice is stated without explanation or is not clearly communicated.

Level 1 The student's response shows a very poor understanding of the requirements of the task. The response addresses some components of the task and the concepts associated with it.

- Computations for area and/or perimeter attempted.
- Little or no understanding of the relationship between area and perimeter shown.

Figure 1 (Cont.)

Holistic Scoring Rubric for Garden Fence Task

PERFORMANCE CRITERIA:

Chance:	understands concepts of chance (certainty and uncertainty, experimentation and theory, probability, dependence and independence)
Data Analysis:	understands concepts of data collection and analysis (population and sampling, central tendency and distribution)
	conducts data analyses (collects data, analyzes central tendency and distribution, displays results in tables, graphs, and charts)
	understands how to interpret data (inference, point of view, uses and misuses)

SCORING

Exemplary

4 points Meets or exceeds all relevant criteria

- shows **extensive** understanding of concepts and procedures both within and beyond the task
- **consistently and purposefully** applies appropriate concepts and procedures

Proficient

3 points Meets all relevant criteria

- shows **thorough** understanding of concepts and procedures required by the task
- **consistently** applies appropriate concepts and procedures

Intermediate

2 points Meets some relevant criteria

- shows **general** understanding of concepts and procedures required by the task
- **generally** applies appropriate concepts and procedures

Novice

1 point Meets few relevant criteria

- shows **rote or partial** understanding of concepts and procedures required by the task
- **occasionally** applies appropriate concepts and procedures

Not Scorable: Attempted with no understanding, off task, not attempted

Figure 2A

Trait Scoring Rubric for Statistics and Probability Applied to Number Cube and School Drop-Out Tasks

Copyright © 1996, Commission on Student Learning, State of Washington, Olympia, WA. All rights reserved. Reproduced by permission.

Communication Performance Criteria

Gathers Information:	plans, obtains information from sources
Interprets Information:	organizes information, clarifies understandings
Represents & Shares Information:	expresses mathematical ideas via physical/pictorial models, tables, charts, graphs, algebraic notation, language; expression appropriate to audience

SCORING

Exemplary
4 points Meets all relevant criteria

- gathers **all applicable** information from appropriate sources
- demonstrates interpretations and understandings in a **clear, systematic**, and **organized** manner
- represents mathematical information and ideas in an **effective** format for the task, situation, and audience

Proficient
3 points Meets most relevant criteria

- gathers **applicable** information from appropriate sources
- demonstrates interpretations and understandings in a **clear** and **organized** manner
- represents mathematical information and ideas in an **expected** format for the task, situation, and audience

Intermediate
2 points Meets some relevant criteria

- gathers information from appropriate sources
- demonstrates interpretations and understandings in an **understandable** manner
- represents mathematical information and ideas in an **acceptable** format for the task, situation, and audience

Novice
1 point Meets few relevant criteria

- gathers **little** information from appropriate sources
- demonstrates interpretations and understandings in a manner that **may be disorganized or difficult to understand**
- represents mathematical information and ideas in a format that **may be inappropriate** for the task, situation, and audience

Not Scorable no attempt, off topic, can't be read

Figure 2B

Trait Scoring Rubric for Mathematical Communication Applied to All Three Tasks

The numbers 1, 1, 1, 2, 3, and 4 are on the first cube and the numbers 1, 2, 2, 3, 4, and 5 are on the second cube.

Each time the sum of the numbers tossed is less than or equal to 5 Alex gets a point. When the sum of the numbers tossed is greater than or equal to 6 Robin gets a point.

2. Use the space below to display, in an organized way, the possible outcomes when Robin and Alex toss the number cubes.

3 points:
- Visual display shows the outcomes two digit combinations (see example).
- All marginal digits and sums are accurate.
- Some indication is given showing the outcomes for which Alex wins versus those for which Robin wins

	1	1	1	2	3	4
1	2	2	2	3	4	5
2	3	3	3	4	5	6
2	3	3	3	4	5	6
3	4	4	4	5	6	7
4	5	5	5	6	7	8
5	6	6	6	7	8	9

☐ = Alex wins

2 points
- Table, chart or visual display gives all or most possible combinations for digits on the cubes.
- Visual display shows that different two digit combinations will yield different sums.
- Chart may not be complete or may have minor errors either in the transfer of the numbers from the cubes to the chart or in the sums resulting from different combinations.

1 point
- Table, chart or visual display gives some possible combinations for digits on the cubes.
- Display is not complete or has several errors either in the transfer of the numbers from the cubes to the chart or in the sums resulting from different combinations.

0 points
- Table, chart or visual display, if given, shows no understanding of how to set up problem.

Figure 3

Item-Level Scoring Rule to be Applied to Item 3 of the Number Cube Task

Appendix A

Middle School Task

The Garden Fence Problem

Today you are going to work on a problem about a garden fence. You will work with a group to investigate ideas related to solving the problem and then work alone to solve the problem yourself. You will be evaluated based on:

- how well you use strategies to solve the problem
- the accuracy of your visual displays
- the accuracy of your computations
- the clarity of your communication about your ideas

You may use a straight edge or ruler and a calculator in your work.

Terms you need to know to do the problem:

area: a measure of the surface of a figure found by multiplying the length times the width of the figure
perimeter: the measurement around the outside of a figure
dimension: the length or the width of a figure

NOTE: Students work in small groups to explore area and perimeter on grid paper. Students look at the relationship between area and perimeter by using a fixed area and looking at rectangles of different perimeters and then by using a fixed perimeter and looking at rectangles of different areas. Once small group work is completed, students work independently.

Read the problem in the box.

Adam, Isaac and Corina have a rectangular vegetable garden that measures 6 ft. by 16 ft. They keep the garden fenced to keep out rodents and neighborhood dogs. They have decided to make the area of their garden larger.

- Adam said they should buy more fencing.
- Isaac said they can fence a larger garden with the fencing they already have.
- Corina said they can fence a larger area using less fencing than they are using for their garden now.

Now, do Numbers 3 through 5.

3. The current garden is 6 feet by 16 feet. What is the current area and perimeter of the garden? Show your work.

4. Tell how each person's idea could result in more garden area than they have right now. Write a sentence for each person telling whether the children could have more garden area using each different idea. Draw pictures and label the dimensions, area, and perimeter to support your statement about each person's idea.

   - Adam's Idea: Add more fence and get more area
   - Isaac's Idea: Use the same fence and get more area
   - Corina's Idea: Use less fence and get more area

5. Although all three ideas may be possible, suppose the children cannot afford to buy more fencing. Choose either Isaac's or Corina's plan. Then write a paragraph describing the results and telling why the plan will result in a larger garden with the same amount of fence or with less fence.

Copyright © 1995, Commission on Student Learning, State of Washington, Olympia, WA. All rights reserved. Reproduced by permission.

Appendix A

High School Task #1

School Drop-Out

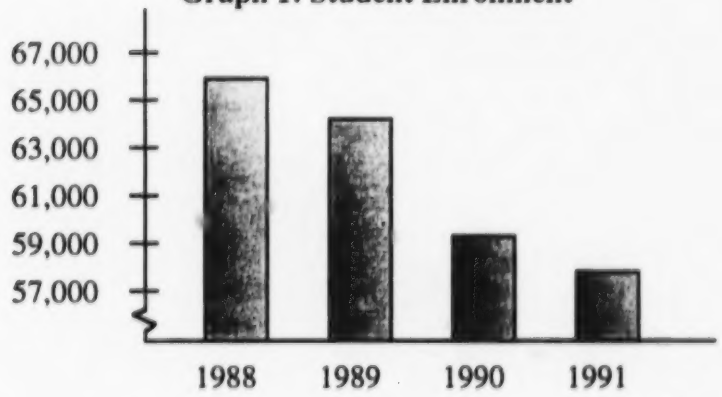
The table below shows the number of high school students and the number of high school dropouts in Washington state for the years 1988 through 1991. The dropout rate for each of these years is also given. In addition, a graph illustrating the data is given for each data set. In the items that follow, you will look at the data in the table and the graphs and then describe the trends in enrollment, number of dropouts, and dropout rate from 1988 to 1991. You will also make predictions of what the numbers probably looked like in 1995 if the trends continued in the same fashion. You will be evaluated based on how well you sow your statistical understandings and how you communicate your mathematical ideas.

Use the table and graphs to do Numbers 1 through 7. You may refer back to the table and graphs as often as you need to.

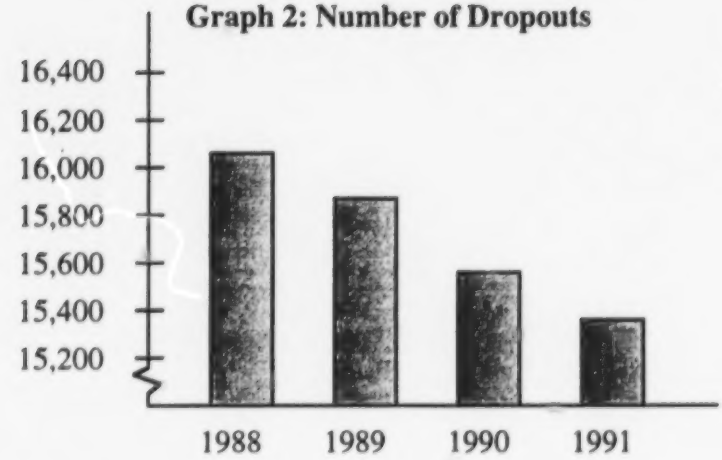
Washington State High School Enrollment Data

	1988	1989	1990	1991
Student Enrollment	65,920	64,260	59,280	57,760
Dropouts	16,060	15,850	15,580	15,390
Dropout Rate	0.24	0.25	0.26	0.27

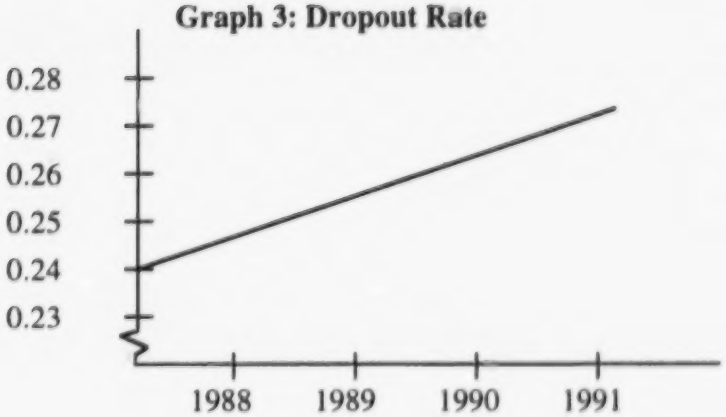
Graph 1: Student Enrollment



Graph 2: Number of Dropouts



Graph 3: Dropout Rate



1. Use the table and Graph 1: Student Enrollment to describe the trend in high school enrollment from 1988 through 1991.

2. Use the table and Graph 2: Number of Dropouts to describe the trend in high school dropouts from 1988 through 1991. Use specific data from the graph or table in your answer.

3. Use the table and Graph 3: Dropout Rate to describe the trend in the Washington high school dropout rate from 1988 through 1991. Use specific data from the graph or table in your answer.

4. Compare the trends in number of dropouts with the dropout rate. Is this situation possible? If it is, explain how it could occur. If it is not, explain why. Refer to data from the table or graphs in your response.

5. Chris argues that the dropout problem is improving. Which graph supports Chris's argument? Tell why.

6. Terry argues that the dropout problem is getting worse. Which graph supports Terry's argument? Tell why.

7. Decide if you think the dropout problem is improving, remaining constant, or getting worse. Draft a brief letter to the Governor. In your letter:
   - state your position about the dropout problem
   - tell what you think 1995 drop-out data looks like given the trends in the data.
   - **use data from the table and graphs to support your position and prediction**
   - add your own knowledge
     - if you think the drop-out problem is getting worse, give one possible cause for the problem
     - if you think the drop-out problem is getting better, give one possible reason for the improvement
     - conclude your letter with a recommendation about how to eliminate school dropout

Copyright © 1995, Commission on Student Learning, State of Washington, Olympia, WA. All rights reserved. Reproduced by permission.

High School Task #2

Number Cube Games

Today you are going to do a mathematics investigation. You will work with a partner to investigate the problem and then write about the investigation. Your work will be evaluated based on whether you:

- complete all steps
- use an appropriate method for displaying data
- display your data in a way that is organized and easily read
- clearly explain your reasoning
- determine the probabilities of each outcome
- use everyday language to express your ideas
- use mathematical notations to express mathematical ideas

Problem to be Investigated

Robin and Alex are playing a game using two number cubes with the numbers 1, 1, 1, 2, 3, and 4 on the first cube and the numbers 1, 2, 2, 3, 4, and 5 on the second cube. They take turns rolling the cubes.

Each time the sum of the numbers tossed is less than or equal to 5 Alex gets a point. When the sum of the numbers tossed is greater than or equal to 6 Robin gets a point.

1. Before investigating further, decide whether you think this is a fair game. Would you rather get Robin's points or Alex's points or does it matter? Write your prediction in the space below.

NOTE: Students then work with a partner to discuss ways to represent the different combinations that can come about with the numbers on the cubes. Once they have discussed ways to represent the data, they then work independently.

2. Use the space below to display, in an organized way, the possible outcomes when Robin and Alex toss the number cubes.

Remember, each time the sum of the numbers tossed is less than or equal to 5 Alex gets a point. When the sum of the numbers tossed is greater than or equal to 6 Robin gets a point.

3. Look at your data display on the previous page. Is this a fair game? Would you rather get Robin's points or Alex's points or does it matter? Explain your reasoning.

4. Write <u>two</u> sentences telling how your conclusion compares to your prediction?

5. Create a new game from what you learned in this investigation.

   If you determined that the original game is a fair game
   - change the numbers on the cubes to make a new game but make sure the game is still fair
   - show the possible outcomes for the new number cubes in an organized display
   - explain why it is still fair

   If you determined that the original game is not a fair game
   - change the numbers on the cubes to make it fair
   - show the possible outcomes for the new number cubes in an organized display
   - explain why it is now a fair game

6. Use the information from the investigation and write a letter to the Double-dealing Game Company that distributes this game. In your letter you are to:
   - Summarize the results of your investigation
   - Discuss the role that probability has in the outcomes of the game
   - Tell them that you have determined that the game is fair or unfair
   - Explain your reasoning
   - Show an organized display to back up your decision and reasoning
   - If the game is fair, present your idea for a new game
   - If the game is not fair, present the changes you made in order to have a fair game
   - Explain your reasoning for the new game and show an organized display to back up your reasoning

Copyright © 1995, Commission on Student Learning, State of Washington, Olympia, WA. All rights reserved. Reproduced by permission.